(19) BUNDESREPUBLIK DEUTSCHLAND

(12) **Offenlegungsschrift**

(10) **DE 41 11 995 A1**

(51) Int. Cl.[5]: G 10 L 7/08

DEUTSCHES PATENTAMT

(21) Aktenzeichen: P 41 11 995.9
(22) Anmeldetag: 12. 4. 91
(43) Offenlegungstag: 15. 10. 92

DE 41 11 995 A1

(71) Anmelder:
Philips Patentverwaltung GmbH, 2000 Hamburg, DE

(72) Erfinder:
Meyer, Peter, Dr., 8510 Fürth, DE; Rühl, Hans-Wilhelm, Dr., 8501 Schwaig, DE

(54) Schaltungsanordnung zur Spracherkennung

(57) Die Erfindung bezieht sich auf eine Schaltungsanordnung zur Spracherkennung. Zur Erkennung eines Sprachsignals ist es nötig, eine Analyse des Sprachsignals mit dem Ziel der Extraktion von charakteristischen Merkmalen vorzunehmen. Die extrahierten Merkmale werden durch sogenannte spektrale Merkmalsvektoren repräsentiert, die mit für das zu erkennende Sprachsignal abgespeicherten Referenz-Merkmalsvektoren verglichen werden. Die Referenz-Merkmalsvektoren werden während einer Trainingsphase, in der ein Sprachsignal mehrmals aufgenommen wird, ermittelt. Das Erkennungsergebnis hängt im wesentlichen von der Güte der spektralen Merkmalsvektoren bzw. Referenz-Merkmalsvektoren ab.
Es wird deshalb vorgeschlagen, eine rekursive Hochpaßfilterung der spektralen Merkmalsvektoren vorzusehen. Hierdurch wird der Einfluß von Störgrößen auf das Erkennungsergebnis vermindert und ein hoher Grad an Sprecherunabhängigkeit der Erkennung erreicht. Dies ermöglicht den Einsatz der Schaltungsanordnung zur Spracherkennung auch in Systemen, die eine sprecherunabhängige Spracherkennung voraussetzen.

DE 41 11 995 A1

Beschreibung

Die Erfindung betrifft eine Schaltungsanordnung zur Spracherkennung mit einer Auswerteschaltung zur Ermittlung von spektralen Merkmalsvektoren von Zeitrahmen eines digitalen Sprachsignals mittels einer Spektralanalyse, zur Logarithmierung der spektralen Merkmalsvektoren und zum Vergleich der logarithmierten spektralen Merkmalsvektoren mit Referenz-Merkmalsvektoren.

Sprecherabhängige Spracherkennungseinrichtungen werden in vielen Bereichen erfolgreich eingesetzt, so z. B. in Systemen, die gesprochenen Text erkennen, verstehen und in eine Handlung umsetzen (akustisch gegebene Befehle zur Steuerung von Geräten), wobei das zu erkennende Sprachsignal oftmals zusätzlich über eine Telefonleitung (Fernwirken über Telefon) übertragen wird.

In dem Buch "Automatische Spracheingabe und Sprachausgabe" von K. Sickert, Haar bei München, Verlag Markt und Technik, 1983, Seiten 223–230 und 322–326 wird der prinzipielle Aufbau einer Spracherkennungseinrichtung beschrieben, bei der das Sprachsignal zunächst analysiert wird, um die informationstragenden Merkmale zu extrahieren. Diese Merkmale werden durch sogenannte Merkmalsvektoren repräsentiert, die mit den in einem Referenzspeicher abgelegten, während einer Lernphase ermittelten Referenz-Merkmalsvektoren in einer Erkennungseinheit verglichen werden.

Aus der Veröffentlichung "Verfahren für Freisprechen, Spracherkennung und Sprachcodierung in der SPS51" von W. Armbrüster, S. Dobler und P. Meyer, PKI Technische Mitteilungen 1/1990, Seiten 35–41 ist eine technische Realisierung einer sprecherabhängigen Spracherkennungseinrichtung bekannt. In dieser Spracherkennungseinrichtung werden bei einer Analyse eines digitalen Sprachsignals der zeitliche Verlauf dieses Signals im Spektralbereich betrachtet und spektrale Merkmalsvektoren ermittelt, die zur Beschreibung der charakteristischen Merkmale des Sprachsignals geeignet sind. In einer Lern- bzw. Trainingsphase, im weiteren als Training bezeichnet, wird jedes zu erkennende Wort mehrmals aufgenommen. Dabei werden jeweils spektrale Merkmalsvektoren ermittelt, woraus durch Mittelung wortspezifische Referenz-Merkmalsvektoren erzeugt werden. Nach Abschluß des Trainings stehen für jedes trainierte Wort Referenz-Merkmalsvektoren, die in einem Referenzmusterspeicher abgelegt sind, zur Verfügung. Im Normalbetrieb, der Testphase, werden für ein zu erkennendes Sprachsignal die spektralen Merkmalsvektoren bestimmt und einer Erkennungseinheit zugeführt, in der ein Vergleich mit den abgespeicherten Referenz-Merkmalsvektoren mittels eines auf der dynamischen Programmierung basierenden Verfahren stattfindet.

Probleme bei der Erzielung eines zuverlässigen Erkennungsergebnisses ergeben sich vor allem durch die Überlagerung des Sprachsignals mit Störgrößen, wie z. B. Verzerrungen des Frequenzganges oder quasistationäre Störgeräusche. Solche Störgrößen werden überwiegend bei der Übertragung des Signals über eine Telefonleitung und/oder durch Hintergrundrauschen bei der Aufnahme eingestreut. Zudem verschlechtern sich die Erkennungsergebnisse, wenn die Ermittlung von Referenz-Merkmalsvektoren während des Trainings unter anderen Aufnahmebedingungen als die Ermittlung von Merkmalsvektoren während der Testphase erfolgt. In diesem Fall kann die Erkennungseinheit den Vergleich

zwischen Merkmalsvektoren und Referenz-Merkmalsvektoren nicht mehr zuverlässig vornehmen, woraus eine Erhöhung der Fehlerrate bei der Erkennung resultiert.

Darüber hinaus werden die Einsatzmöglichkeiten von Spracherkennungseinrichtungen vor allem dadurch eingeengt, daß die Mehrzahl der bisherigen technischen Realisierungen lediglich zur sprecherabhängigen Spracherkennung, die ein Training durch den jeweiligen Benutzer impliziert, geeignet sind. Ein Einsatz von solchen sprecherabhängigen Spracherkennungseinrichtungen in Systemen, in denen der gesprochene Text von häufig wechselnden Benutzern erkannt und/oder beantwortet werden soll (z. B. vollautomatische Auskunftssysteme mit sprachlichem Dialog) ist nur schlecht möglich.

Die Aufgabe der vorliegenden Erfindung ist es deshalb, eine Schaltungsanordnung zur Spracherkennung der eingangs genannten Art so zu verbessern, daß eine sprecherunabhängige Erkennung ermöglicht und der Einfluß von Störgrößen auf das Erkennungsergebnis vermindert wird.

Die Aufgabe wird erfindungsgemäß dadurch gelöst, daß vor dem Vergleich mit den Referenz-Merkmalsvektoren in der Auswerteschaltung eine rekursive Hochpaßfilterung der spektralen Merkmalsvektoren vorgesehen ist.

Die spektralen Merkmalsvektoren enthalten eine Zahl von Komponenten, die während einer Merkmalsextraktion durch zahlreiche Verarbeitungsschritte ermittelt werden. Hierbei werden die Komponenten unter anderem einer Logarithmierung unterworfen. Stationäre oder langsam veränderliche Störungen bzw. Änderungen des Frequenzganges, die während der Aufnahme oder der Übertragung des Sprachsignals dem zu erkennenden Sprachsignal überlagert wurden, führen in den logarithmierten Komponenten der Merkmalsvektoren zu additiven Termen, die durch eine Hochpaßfilterung der Komponenten der spektralen Merkmalsvektoren unterdrückt werden. Daneben wird durch den Einsatz einer rekursiven Hochpaßfilterung eine erhebliche Verbesserung der Sprecherunabhängigkeit der Spracherkennung erzielt. Die Schaltungsanordnung zur Spracherkennung muß im Normalfall nur noch einmal trainiert werden und ist anschließend in der Lage, Sprachsignale auch dann zu erkennen, wenn sie von Personen gesprochen werden, die die Schaltungsanordnung zur Spracherkennung nicht trainiert haben. Hierdurch wird ein enormes Anwendungssprektrum für die erfindungsgemäße Schaltungsanordnung zur Spracherkennung eröffnet. Sie kann z. B. zur Realisierung eines Telefonauskunftssystems mit sprachlichem Dialog oder zur Steuerung von Geräten mittels Spracheingabe eingesetzt werden, wobei das Training der Schaltungsanordnung zur Spracherkennung bereits vom Hersteller vorgenommen werden kann und somit ein Trainieren durch den Benutzer entfällt. Darüber hinaus bewirkt die vor dem Vergleich mit Referenz-Merkmalsvektoren vorgesehene Filterung der spektralen Merkmalsvektoren mit einem rekursiven Hochpaß — selbstverständlich werden auch die zur Bestimmung der Referenz-Merkmalsvektoren ermittelten spektralen Merkmalsvektoren während der Trainigsphase dieser Filterung unterworfen — eine deutliche Reduzierung des Einflusses von stationären Störgeräuschen (z. B. durch Brummen in einer Telefonverbindung hervorgerufen) und eine verbesserte Unterdrückung von Frequenzgangverzerrungen. Es sei an dieser Stelle bemerkt, daß die Auswerteschaltung der Schaltungsanordnung zur Spracherken-

nung wahlweise durch einen Prozessor oder durch diskrete Bauelemente gebildet wird. Darüber hinaus können ein oder mehrere, der in der Auswerteschaltung vorgesehenen Schritte wahlweise mit diskreten Bauelementen oder als Rechnerprogramm eines Prozessors realisiert werden.

In einer vorteilhaften Ausgestaltung der Erfindung wird vorgeschlagen, die rekursive Hochpaßfilterung von der Auswerteschaltung durch Berechnung hochpaßgefilterter spektraler Merkmalsvektoren M(n, i) gemäß der Vorschrift

$$M(n,i) = V(n,i) - V(n-1,i) + C \cdot M(n-1,i)$$

vorzunehmen, wobei n einen Zeitrahmen, V(n, i) die ungefilterten spektralen Merkmalsvektoren des Zeitrahmens n, M(n−1, i) die spektralen Merkmalsvektoren des Zeitrahmens n−1, i eine spektrale Komponente eines spektralen Merkmalsvektors M bzw. V und C eine vorgegebene Konstante bezeichnet. Bei einer Untersuchung von mehreren verschiedenen rekursiven und nicht rekursiven Hochpaßfilterungen hat sich gezeigt, daß die vorgeschlagene rekursive Hochpaßfilterung erster Ordnung zu den besten Erkennungsergebnissen führt. Die Güte dieser Erkennungsergebnisse hängt zudem im hohen Maße von dem für die Konstante C gewählten Wert ab. Für die Konstante C muß ein Wert im Bereich von $0 < C < 1$ gewählt werden. Da für einen Wert von $C = 0$ der rekursive Hochpaß zu einem Differenzierer entartet und für einen Wert $C = 1$ nur ein Gleichanteil der Komponenten des spektralen Merkmalsvektors unterdrückt wird, hat sich für C ein Wert von ungefähr 0,7 als vorteilhaft erwiesen, um sprecherspezifische Merkmale in den spektralen Merkmalsvektoren zu unterdrücken. Bei zu großen Abweichungen von diesem Wert verschlechtern sich die Erkennungsergebnisse deutlich.

In einer Ausgestaltung der Erfindung ist für eine in der Auswerteschaltung vor zunehmende Spektralanalyse eine Aufteilung des digitalen Sprachsignals in sich überlappende Zeitrahmen, eine nachfolgende Wichtung der Abtastwerte eines Zeitrahmens mit einem Hamming-Fenster und eine schnelle Fouriertransformation mit einer anschließenden Betragsbildung zur Ermittlung von spektralen Merkmalsvektoren vorgesehen.

Im einzelnen bedeutet dies, daß jeweils eine bestimmte Zahl von Abtastwerten des digitalen Sprachsignals zu Blöcken, im weiteren als Zeitrahmen bezeichnet, zusammengefaßt wird. Jeder Abtastwert ist dabei in mehreren Zeitrahmen enthalten, d. h. die Zeitrahmen sind zeitlich versetzt und überlappen sich. Die Abtastwerte eines Zeitrahmens bilden die Grundlage für die Ermittlung eines dem Zeitrahmen zugeordneten spektralen Merkmalsvektors. Bei der Bestimmung des spektralen Merkmalsvektors werden die Abtastwerte eines Zeitrahmens mit einem Hamming-Fenster gewichtet, wie es z. B. in dem Buch "Automatische Spracheingabe und Sprachausgabe" von K. Sickert, Haar bei München, Verlag Markt und Technik, 1983, Seiten 118−119 beschrieben ist. Die Abtastwerte jedes Zeitrahmens werden im Anschluß daran einer schnellen Fourier-Transformation (FFT) unterworfen und aus dem resultierenden Spektrum wird durch eine Quadrierung und eine Betragsbildung das Leistungsdichtespektrum ermittelt. Die spektralen Werte des Leistungsdichtespektrums eines Zeitrahmens bilden die Komponenten des zugeordneten Merkmalsvektors.

Es sei hier erwähnt, daß die Bestimmung der spektralen Merkmalsvektoren alternativ durch eine Filterbank-Analyse, wie sie aus dem Buch "Automatische Spracheingabe und Sprachausgabe" von K. Sickert, Haar bei München, Verlag Markt und Technik, 1983, Seiten 129−131 bzw. Seite 324 bekannt ist, vorgenommen werden kann. Die in der Erfindung eingesetzte, auf der schnellen Fouriertransformation basierende Spektralanalyse bietet den Vorteil, daß sie im Gegensatz zur Filterbank-Analyse, auch mittels eines Programms in einem Prozessor, z. B. in einem Signalprozessor, realisierbar ist.

In einer weiteren vorteilhaften Ausgestaltung der Erfindung ist die Auswerteschaltung zur Reduzierung von Komponenten der spektralen Merkmalsvektoren durch eine Faltung mit Faltungskernen eingerichtet. Die Faltungskerne (Mittenfrequenzen) werden so gewählt, daß sie in regelmäßigen Abständen auf der sogenannten "mel"-Skala (Melodie-Kurve) der subjektiven musikalischen Tonhöhe liegen, wodurch eine Auswahl von Komponenten der spektralen Merkmalsvektoren nach psycho-akustischen Aspekten gewährleistet ist. Der Verlauf der "mel"-Skala ist z. B. aus dem Buch "Das Ohr als Nachrichtenempfänger" von E. Zwicker und R. Feldtkeller, S. Hirzel Verlag, Stuttgart, 1967 bekannt.

Die aus der Faltung resultierende Unterabtastung führt in vorteilhafter Weise zu einer erheblichen Reduzierung der Komponenten der spektralen Merkmalsvektoren und damit zu einer deutlichen Datenreduktion.

Eine weitere Ausführungsform zeichnet sich dadurch aus, daß eine in der Auswerteschaltung vorzunehmende Logarithmierung der spektralen Merkmalsvektoren bei einer auf der schnellen Fouriertransformation basierenden Spektralanalyse vor der Faltung vorgesehen ist. Hierdurch wird eine Kompandierung der Komponenten der spektralen Merkmalsvektoren erreicht, woraus eine erhebliche Reduktion der zu verarbeitenden Datenmenge resultiert.

Eine Verringerung des Einflusses von Störgrößen, die von im allgemeinen unbekannten Eigenschaften eines Übertragungsweges des Sprachsignals abhängig sind, wird in einer Ausgestaltung durch eine Intensitätsnormierung der spektralen Merkmalsvektoren erzielt. Es wird hierzu ein Mittelwert der Komponenten eines jeden spektralen Merkmalsvektors berechnet und anschließend von jeder Komponente subtrahiert. Der Mittelwert entspricht einer mittleren Energie eines spektralen Merkmalsvektors und wird deshalb als weitere Komponente eines spektralen Merkmalsvektors aufgenommen. Durch die vorgeschlagene Intensitätsnormierung wird zudem die für die Erkennung nachteilige Abhängigkeit der Komponenten von der Lautstärke des Sprachsignals annähernd beseitigt und die Leistungsfähigkeit der Spracherkennung verbessert.

Im folgenden soll anhand des in den **Fig.** 1 bis 3 schematisch dargestellten Ausführungsbeispiels die Erfindung näher beschrieben und erläutert werden.

Es zeigt:

Fig. 1 ein Blockschaltbild einer Schaltungsanordnung zur Spracherkennung,

Fig. 2 ein Flußablaufdiagramm der Spracherkennung, wie sie in der Auswerteschaltung der Schaltungsanordnung zur Spracherkennung vorgesehen ist.

Fig. 3 ein Flußablaufdiagramm der Merkmalsextraktion, wie sie in der Auswerteschaltung der Schaltungsanordnung zur Spracherkennung vorgesehen ist.

Fig. 1 zeigt den Aufbau einer Schaltungsanordnung zur Spracherkennung. Ein zu erkennendes analoges Sprachsignal 1, das beispielsweise über ein Mikrofon

oder eine Telefonleitung zugeführt wird und beispielsweise im Frequenzbereich von 0,3 bis 3,4 kHz liegt, wird durch einen Analog-Digital-Wandler 2 mit einer Frequenz von 8 kHz abgetastet und in ein digitales Sprachsignal 3 umgewandelt. Eine Auswerteschaltung 4, die im Ausführungsbeispiel durch einen Signalprozessor mit einem Speicher realisiert ist, ermittelt aus dem digitalen Sprachsignal 3 ein Erkennungssignal 5, welches in einem geeigneten Datenformat Informationen über die im digitalen Sprachsignal 3 erkannten Wörter enthält. Das Erkennungssignal 5 bildet die Grundlage für eine weiterführende Verarbeitung, wie z. B. der Auslösung von Handlungen (Steuerung von Geräten) oder der Ausgabe einer Antwort durch eine Spracherzeugung (Dialogauskunftssystem). Die Schaltungsanordnung kann selbstverständlich in alle gängigen Systeme eingebracht werden, in denen die Erkennung von einzelnen Wörtern oder eine kontinuierliche Spracherkennung vorgesehen ist. Eine Auflistung von Anwendungsmöglichkeiten einer Schaltungsanordnung zur Spracherkennung ist z. B. in dem Buch "Sprachverarbeitung und Sprachübertragung" von K. Fellbaum, Berlin, Springer Verlag, 1984, Seite 204 zu finden.

Fig. 2 verdeutlicht anhand eines Flußdiagramms die in der Auswerteschaltung 4 von **Fig. 1** vorgesehenen Schritte zur Erzeugung des Erkennungssignals 5. Dabei sind gleiche Teile mit den gleichen Bezugszeichen versehen. Die Auswerteschaltung 4 wird im Ausführungsbeispiel durch einen Signalprozessor mit einem Speicher gebildet, der entsprechend den Flußablaufdiagrammen von **Fig. 2** und **Fig. 3** programmiert ist. Aus dem digitalen Sprachsignal 3 werden mit Hilfe einer Merkmalsextraktion (Block 20), deren Schritte in **Fig. 3** detailliert beschrieben werden, spektrale Merkmalsvektoren 21 gewonnen.

In der Auswerteschaltung 4 werden durch eine Verzweigung 22 die zwei Betriebsarten "Training" und "Testphase" unterschieden. Bevor eine Erkennung von Wörtern des digitalen Sprachsignals 3 möglich ist, muß die Schaltungsanordnung während des Trainings zunächst mit den Worten trainiert werden, die später während der Testphase erkannt werden sollen. Während des Trainings wird jedes zu trainierende Wort mehrmals aufgenommen und der Schaltungsanordnung zugeführt. Bei jedem Aufnahmevorgang wird eine Merkmalsextraktion (Block 20) vorgenommen und die resultierenden, für das trainierte Wort spezifischen spektralen Merkmalsvektoren 21 einem Trainingsblock (Block 23) zugeführt. Im Trainingsblock (Block 23) werden aus den, aus mehreren Aufnahmen stammenden Merkmalsvektoren in bekannter Weise eine Reihe von wortspezifischen Referenz-Merkmalsvektoren gebildet, die anschließend abgespeichert (Block 24) werden. Nach Abschluß des Trainings beinhaltet der Speicher für jedes trainierte Wort Referenz-Merkmalsvektoren, auf die während einer Erkennung (Block 25) in der Testphase zugegriffen wird.

In der Testphase wird, wie im Training, für das zu erkennende digitale Sprachsignal 3 eine Merkmalsextraktion (Block 20) vorgenommen. Die resultierenden spektralen Merkmalsvektoren 21 werden nun jedoch über die Verzweigung 22 der Erkennung (Block 25) zugeführt. Die Erkennung (Block 25) führt einen Vergleich der spektralen Merkmalsvektoren 21 mit den abgespeicherten (Block 24) Referenz-Merkmalsvektoren durch und liefert ein Erkennungssignal 5, daß das Erkennungsergebnis in geeigneter Form wiedergibt und das Ausgangssignal der Schaltungsanordnung zur Spracherkennung darstellt.

Aufbau, Abläufe bzw. Funktionsweise des Trainingsblocks (Block 23), die Abspeicherung der Referenz-Merkmalsvektoren (Block 24) sowie der Erkennung (Block 25) sind bekannt aus der Veröffentlichung "Verfahren für Freisprechen, Spracherkennung und Sprachcodierung in der SPS51" von W. Armbrüster, S. Dobler und P. Meyer, PKI Technische Mitteilungen 1/1990, Seiten 35–41 und/oder aus den Druckschriften "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition" von H. Mey IEEE Transactions ASSP, Vol. ASSP-32, No. 2, April 1984, Seiten 263–271 und "Speaker-dependent connectedspeech recognition via dynamic programming and statistical methods" von H. Boulard et al., in K. Kohler, Bibliotheca Phonetical, (Karger, Basel), No.12, 1985, Seiten 115–148.

Fig. 3 zeigt ein Flußablaufdiagramm der Merkmalsextraktion, wie sie in der Auswerteschaltung der Schaltungsanordnung zur Spracherkennung vorgesehen ist. In **Fig. 3** werden anhand eines Flußablaufdiagramms die notwendigen Schritte für eine Merkmalsextraktion (vgl. Block 20, **Fig. 2**), wie sie in der Auswerteschaltung 4 der Schaltungsanordnung 6 zur Spracherkennung vorgesehen ist, näher erläutert.

Das digitale Sprachsignal 3 wird zunächst in sich überlappende, aus jeweils 256 Abtastwerten bestehende Zeitrahmen gemäß der Vorschrift (Block 30):

$$B(n, s) = \{s(96 \cdot n), \ldots, s(96 \cdot n + 255)\}$$

aufgeteilt, wobei n einen Zeitrahmen, s die Abtastwerte des digitalen Sprachsignals 3 und B(n, s) die 256 zu einem Zeitrahmen n gehörenden Abtastwerte s bezeichnet. Die Vorschrift besagt, daß jeder Zeitrahmen n aus 256 aufeinanderfolgenden Abtastwerten s des digitalen Sprachsignals 3 besteht, wobei jeweils nach 96 Abtastwerten ein neuer Zeitrahmen gebildet wird, so daß sich die Zeitrahmen überlappen. Da die Abtastung mit einer Rate von 8 kHz erfolgt, wird alle 96/8000 s = 12 ms ein neuer Zeitrahmen gebildet, der 256 Abtastwerte enthält. Wie Block 31 zeigt, wird jeder Zeitrahmen anschließend einer Wichtung mit einem Hamming-Fenster unterworfen wie es z. B. aus dem Buch "Automatische Spracheingabe und Sprachausgabe" von K. Sickert, Haar bei München, Verlag Markt und Technik, 1983, Seite 119, bekannt ist. Dazu wird eine Multiplikation mit einem Vektor H, der die Koeffizienten des Hamming-Fensters enthält, gemäß

$$B(n, s) = B(n, s) \cdot H$$

vorgenommen. Nach der Wichtung mit dem Hamming-Fenster (Block 31) wird für jeden Zeitrahmen n ein logarithmiertes Leistungsdichtespektrum ermittelt (Block 32), in dem durch eine schnelle Fouriertransformation (FFT) ein komplexes Spektrum des Zeitrahmens n berechnet und daraus durch Bildung eines Betragsquadrates die Leistungsdichtespektren B(n, f), wobei f die Frequenz bezeichnet, ermittelt werden. Durch Logarithmierung der Leistungsdichtespektren B(n, f) resultieren die logarithmierten Leistungsdichtespektren B(n, f) der Zeitrahmen. Die logarithmierten Leistungsdichtespektren B(n, f) werden somit gemäß der Vorschrift

$$B(n, f) = \log(|FFT(B(n, s))|^2)$$

ermittelt, wobei B(n, s) die mit dem Hamming-Fenster

gewichteten Abtastwerte eines Zeitrahmens n und FFT symbolisch die schnelle Fouriertransformation bezeichnet. Eine solche Bestimmung der logarithmierten Leistungsdichtespektren der Zeitrahmen ist z. B. aus der Veröffentlichung "Verfahren für Freisprechen, Spracherkennung und Sprachcodierung in der SPS51" von W. Armbrüster, S. Dobler und P. Meyer, PKI Technische Mitteilungen 1/1990, Seiten 35–41 bekannt.

Die resultierenden logarithmierten Leistungsdichtespektren B(n, f) der Zeitrahmen enthalten jeweils 256 spektrale Werte. Durch Faltung der logarithmierten Leistungsdichtespektren der Zeitrahmen mit 15 Faltungskernen K(f, i) gemäß

$$V(n,i) = B(n,f) \cdot K(f,i) \text{ mit } i = 1,\ldots,15$$

wobei V(n, i) einen spektralen Merkmalsvektor, n den Zeitrahmen, "·" das Symbol für die Faltungsoperation und i eine Komponente des spektralen Merkmalsvektors V(n, i) bezeichnet, erhält man für jeden Zeitrahmen n einen spektralen Merkmalsvektor V(n, i). Die Faltungskerne sind, wie dies bereits in der Beschreibungseinleitung beschrieben wurde, so gewählt, daß sie bei der Faltung fünfzehn auf der "mel"-Skala gleichverteilte Spektralwerte aus den Leistungsdichtespektren der Zeitrahmen extrahiert werden, die zusammen die Komponenten eines spektralen Merkmalsvektors V(n, i) bilden. Die in Block 33 vorgenommene Faltung und die anschließende Zusammenfassung der resultierenden Komponenten zu einem spektralen Merkmalsvektor V(n, i) führt zu einer erheblichen Datenreduktion und vereinfacht die spätere Erkennung.

Wie Block 34 zeigt, wird für jeden spektralen Merkmalsvektor V(n, i) die mittlere Energie V(n, 0) gemäß

$$V(n,0) = \Sigma V(n,i)/15, i = 1,\ldots,15$$

bestimmt und als Komponente i = 0 in den spektralen Merkmalsvektor V(n, i) aufgenommen. Zudem wird die mittlere Energie V(n, 0) von jeder Komponente i = 1, .., 15 eines spektralen Merkmalsvektors subtrahiert. Dies entspricht einer Normierung der spektralen Merkmalsvektoren auf die mittlere Energie gemäß der Vorschrift:

$$V(n,i) = V(n,i) - V(n,0), i = 1,\ldots,15$$

Es resultiert für jeden Zeitrahmen ein aus 16 Komponenten bestehender spektraler Merkmalsvektor V(n, i).

Anschließend ist für die spektralen Merkmalsvektoren V(n, i) eine rekursive Hochpaßfilterung gemäß der Vorschrift

$$M(n,i) = V(n,i) - V(n-1,i) + C \cdot M(n-1,i)$$

vorgesehen, wobei M(n, i) die hochpaßgefilterten spektralen Merkmalsvektoren, V(n−1, i) die spektralen Merkmalsvektoren des Zeitrahmens n−1, M(n−1, i) die hochpaßgefilterten spektralen Merkmalsvektoren des Zeitrahmens n-1, und C eine vorgegebene Konstante bezeichnet. Für die Konstante C wird ein Wert von ungefähr 0,7 gewählt. Die resultierenden spektralen Merkmalsvektoren M(n, i) bilden die der weiteren Verarbeitung zugrundeliegenden Merkmalsvektoren 21.

Patentansprüche

1. Schaltungsanordnung (6) zur Spracherkennung mit einer Auswerteschaltung (4) zur Ermittlung von spektralen Merkmals-Vektoren (21) von Zeitrahmen eines digitalen Sprachsignals (3) mittels einer Spektralanalyse, zur Logarithmierung (32) der spektralen Merkmalsvektoren (21) und zum Vergleich der logarithmierten spektralen Merkmalsvektoren (21) mit Referenz-Merkmalsvektoren (26), **dadurch gekennzeichnet**, daß vor dem Vergleich mit den Referenz-Merkmalsvektoren (26) in der Auswerteschaltung (4) eine rekursive Hochpaßfilterung (35) der spektralen Merkmalsvektoren (21) vorgesehen ist.

2. Schaltungsanordnung nach Anspruch 1, dadurch gekennzeichnet, daß die rekursive Hochpaßfilterung (35) von der Auswerteschaltung (4) durch Berechnung hochpaßgefilterter spektraler Merkmalsvektoren M(n, i) gemäß der Vorschrift

$$M(n,i) = V(n,i) - V(n-1,i) + C \cdot M(n-1,i)$$

vorgenommen wird, wobei V(n, i) die ungefilterten spektralen Merkmalsvektoren, n einen Zeitrahmen, i eine spektrale Komponente eines spektralen Merkmalsvektors M bzw. V und C eine vorgegebene Konstante bezeichnet.

3. Schaltungsanordnung nach Anspruch 2, dadurch gekennzeichnet, daß für die Konstante C ein Wert im Bereich von $0 < C < 1$ gewählt wird.

4. Schaltungsanordnung nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß die in der Auswerteschaltung (4) vorzunehmende Spektralanalyse eine Aufteilung (30) des digitalen Sprachsignals (3) in sich überlappende Zeitrahmen, eine nachfolgende Wichtung (31) der Abtastwerte eines Zeitrahmens mit einem Hamming-Fenster, eine schnelle Fouriertransformation (32) mit einer anschließenden Betragsbildung zur Ermittlung von spektralen Merkmalsvektoren (21) vorsieht.

5. Schaltungsanordnung nach Anspruch 4, dadurch gekennzeichnet, daß die Auswerteschaltung (4) zur Reduzierung der Zahl von Komponenten der spektralen Merkmalsvektoren (21) durch eine Faltung (33) mit Faltungskernen eingerichtet ist.

6. Schaltungsanordnung nach Anspruch 5, dadurch gekennzeichnet, daß in der Auswerteschaltung (4) die Logarithmierung (32) der spektralen Merkmalsvektoren (21) bei einer auf der schnellen Fouriertransformation basierenden Spektralanalyse vor der Faltung (33) vorgesehen ist.

7. Schaltungsanordnung nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, daß die Auswerteschaltung (4) vor der rekursiven Hochpaßfilterung (35) zur Intensitätsnormierung (34) der spektralen Merkmalsvektoren (21) bestimmt ist.

Hierzu 2 Seite(n) Zeichnungen

— Leerseite —

Nummer: DE 41 11 995 A1
Int. Cl.[5]: G 10 L 7/08
Offenlegungstag: 15. Oktober 1992

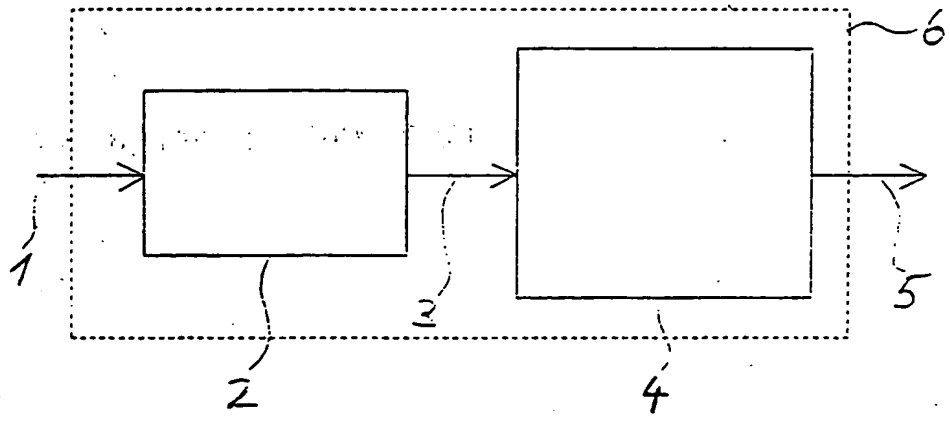

FIG. 1

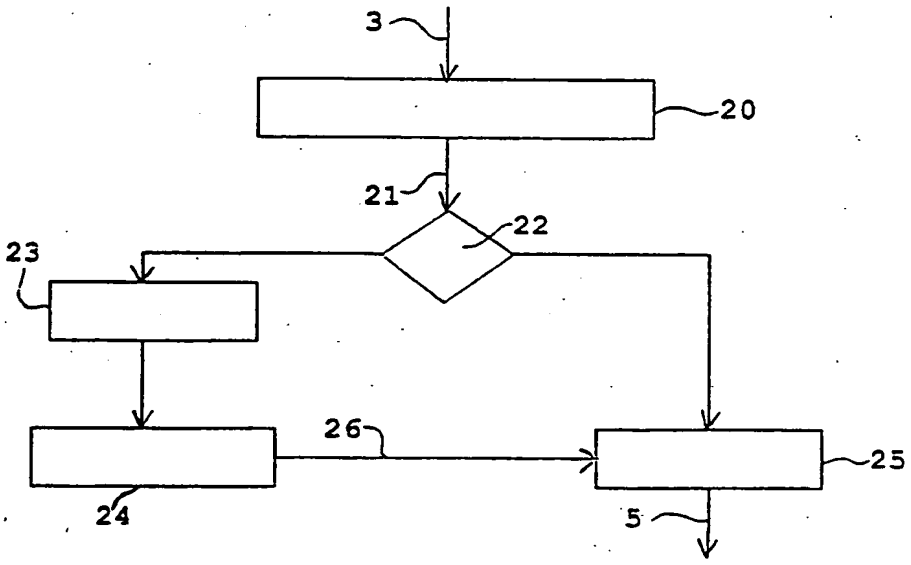

FIG. 2

Nummer: DE 41 11 995 A1
Int. Cl.[5]: G 10 L 7/08
Offenlegungstag: 15. Oktober 1992

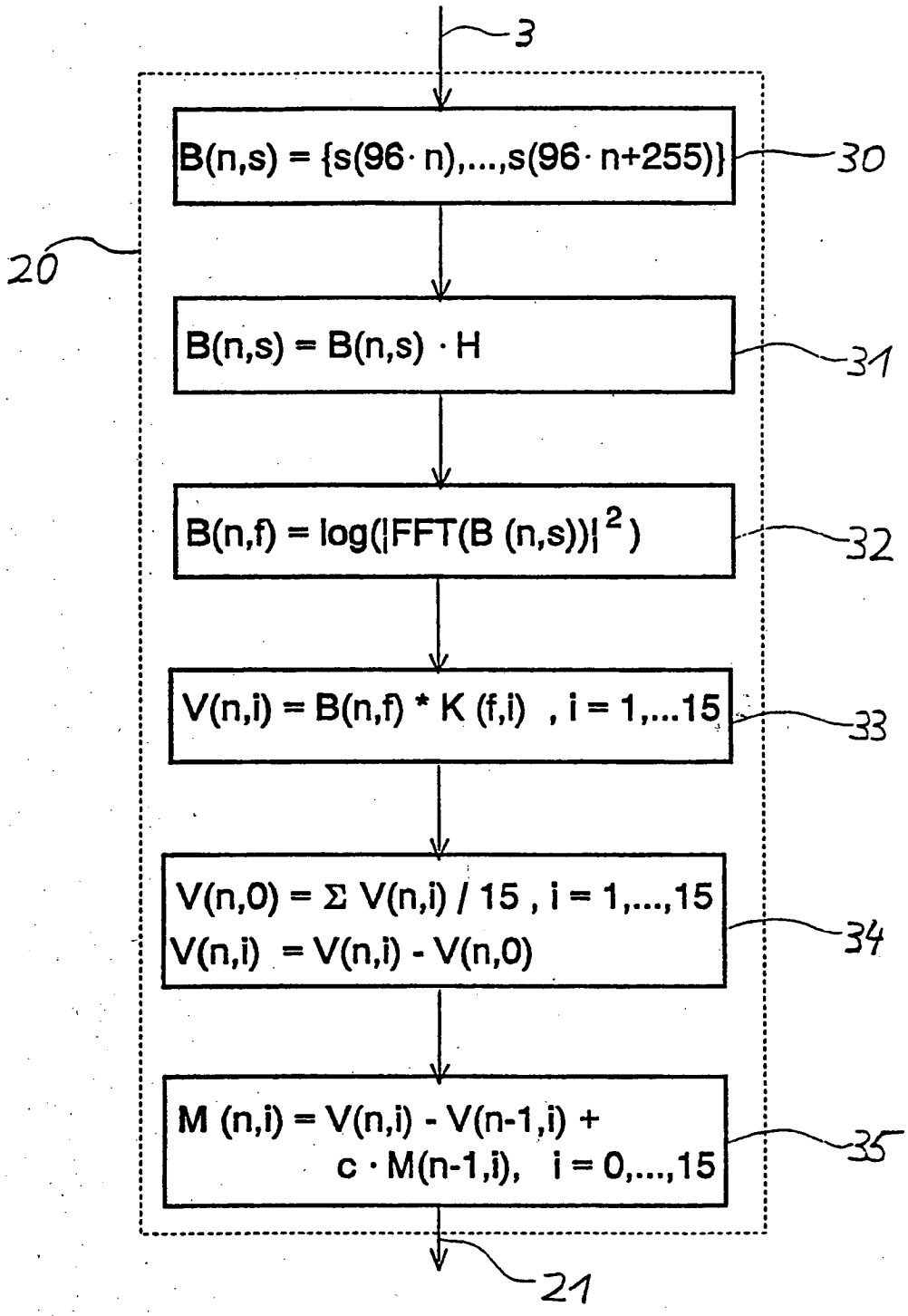

FIG.3

Circuit for speech recognition.

Publication number: DE4111995
Publication date: 1992-10-15
Inventor: MEYER PETER DR (DE); RUEHL HANS-WILHELM DR (DE)
Applicant: PHILIPS PATENTVERWALTUNG (DE)
Classification:
- international: *G10L15/06;* G10L15/00; (IPC1-7): G10L7/08
- european: G10L15/10; G10L15/28H
Application number: DE19914111995 19910412
Priority number(s): DE19914111995 19910412

Also published as:
EP0508547 (A:
JP5108099 (A)
EP0508547 (A:
EP0508547 (B:

Abstract not available for DE4111995
Abstract of corresponding document: **EP0508547**

To recognise a speech signal, it is necessary to carry out an analysis of the speech signal with the aim o extracting characteristic features. The extracted features are represented by so-called spectral feature vectors which are compared with the reference feature vectors stored for the speech signal to be recognised. The reference feature vectors are determined during a training phase in which a speech signal is recorded several times. The result of the recognition essentially depends on the quality of the spectral feature vectors or reference feature vectors. It is therefore proposed to provide recursive high-pass filtering of the spectral feature vectors. This reduces the influence of interference variables on the result of the recognition and provides for a high degree of speaker-independence of the recognition. This makes it possible to use the circuit arrangement for speech recognition even in systems which presuppose speaker-independent speech recognition.

THIS PAGE BLANK (USPTO)

Docket # 2004P00324
Applic. #
Applicant: T. Fingscheidt, et al.
Lerner Greenberg Stemer LLP
Post Office Box 2480
Hollywood, FL 33022-2480
Tel: (954) 925-1100 Fax: (954) 925-1101